立地行動における相互動学的意思決定とその解空間の性質

増田　慧樹[1]・羽藤　英二[2]

[1]学生会員　東京大学　工学系研究科社会基盤学専攻（〒113–8656　東京都文京区本郷 7-3-1）
E-mail: masuda@bin.t.u-tokyo.ac.jp
[2]正会員　東京大学大学院教授　工学系研究科社会基盤学専攻（〒113–8656　東京都文京区本郷 7-3-1）
E-mail: hato@bin.t.u-tokyo.ac.jp

社会基盤整備と利用者の行動は互いに影響し合うため，複数主体の相互動学的な意思決定の分析が重要である．本研究では，計画者と利用者の動学的意思決定を記述する Stackelberg ゲームを多期間に拡張し，両者の読み合いを表現した．展開形ゲームを用いて，計画者の社会基盤整備と住民の居住地選択の相互動学的な関係をモデル化し，均衡解のパターンを整理した．

また，大規模なゲームにおいては，状態空間の爆発により均衡解の計算が困難になる．この課題に対して，マルチエージェント深層強化学習の AlphaZero を導入し，各状態における最適反応を計算した．その結果，最適反応の学習後は，計画者・住民ともに戦略が変化し，利得が増加した．複数主体の相互作用を先読みして，現在の最適な行動を決定する手法の有効性を示した．

Key Words: *location choice, social interaction, extensive-form game, multi-agent reinforcement learning*

1. はじめに

社会基盤の整備計画と利用者の行動は，互いに依存し合っており，効果的な社会基盤整備を行うためには，その相互動学的な関係の分析が重要である．例えば，本研究で扱う災害リスクエリアの立地行動において，計画者は，道路や公共施設等のインフラを整備した場合の住民の居住地移転を予想して，防災的観点からインフラ整備場所を決定する．これに対して，住民もまた，自らの居住地選択の結果，計画者が対応するインフラ整備を行うと予想して居住地を決定する．そのため，災害リスクの高い場所に住民が先に立地し，それを後追いするように行政が防災インフラを整備する，という現象が起こり得る．このような複数主体の再帰的な意思決定は，chicken-or-egg dilemma と呼ばれ，道路整備と旅行者の経路選択[1)] や，電気自動車の充電設備と普及率[2)] などの事例が研究されてきた．

複数主体の相互動学的な意思決定を分析するため，既往研究では，Stackelberg ゲームが用いられてきた．Stackelberg ゲームでは，計画者を leader，利用者を follower とし，follower が leader の選択に対して最適に反応することを前提に，leader が自らの行動を決定する．交通計画の分野では，朝倉[1)] は旅行者 (follower) の利用者均衡を制約として，計画者 (leader) の最適ネットワーク計画を二段階最適化問題として定式化している．近年では，警備側 (leader) の警備計画を観察した攻撃者 (follower) が悪意を持って攻撃してくる場合の，警備人員の最適な配置計画を決定する，security game の研究[3)4)] が盛んである．しかし，Stackelberg ゲームでは leader と follower の一回きりの行動しか記述していない．follower が自らの行動の後の leader の行動を予想した上で，行動を決定する枠組みにはなっておらず，完全に相互動学的な意思決定が記述できているわけではない．そのため，Stackelberg ゲームにおける，複数主体の互いの行動の読み合いが多期間に渡る場合，どのような均衡が現れるかは明らかではない．

そこで，本研究では，Stackelberg ゲームを多期間に拡張し，展開形ゲームとして記述することにより，立地行動における複数主体の相互動学的な意思決定を分析する．特に，災害リスクエリアにおける，計画者の社会基盤整備と住民の居住地選択を対象に，均衡解の種類と出現条件について示す．また，主体の行動選択肢や期間数が増加すると，ゲームの状態空間の爆発により，全探索で解を見つけることは困

難になる．これに対して，マルチエージェント深層強化学習を用いて，各状態におけるプレーヤーの最適戦略を計算することを試みる．

本論文の構成を示す．1 章では，研究の背景を示し，相互動学的な意思決定の分析の枠組みを述べた．2 章では，複数主体の動学ゲームに関する既往研究を整理し，本研究の焦点を述べる．3 章では，災害リスクエリアにおける計画者の社会基盤整備と住民の居住地選択を展開形ゲームで記述し，均衡概念を示す．これに基づき，4 章では展開形ゲームの解空間を分析する．5 章では，大規模ゲームに対応するため，マルチエージェント深層強化学習を導入する．最後に，6 章で本研究の成果をまとめ，今後の展望を述べる．

2. 既往研究の整理

(1) 交通計画における Stackelberg ゲーム

複数主体の相互動学的な意思決定を記述する手法として，Stackelberg ゲームを用いた，交通計画分野の研究の蓄積がある．

朝倉[1] は，旅行者の OD 需要の変化を内生化した道路ネットワークデザイン問題を定式化するため，Stackelberg ゲームを用いた．計画者を leader，旅行者を follower としたとき，計画者の道路交通網計画に応じて，旅行者は経路選択行動を最適に変化させるため，それを先読みして計画者は計画を立てる必要がある．これは，旅行者の利用者均衡を制約としてもつ計画者の道路投資最適化問題となり，二段階最適化問題として定式化し，解法を提示した．

Bell ら[5] は，Stackelberg ゲームを道路ネットワークの信頼性分析に応用した．彼らは，悪意を持って道路ネットワークを攻撃する主体を demon (leader) とし，demon の攻撃を観察した上で経路選択を行う旅行者 (follower) との Stackelberg 均衡を考えた．均衡において demon の選択するリンクは，旅行者の経路選択を先読みした上で旅行者に最大の総旅行時間を課すリンクで，脆弱なリンクと言える．また均衡において旅行者が選択する経路は，災害や事故などの突発的なリスクに対して悲観的な旅行者の経路選択行動と解釈できる．Stackelberg ゲームは，各主体の完全情報と合理的な先読み行動を仮定しているため，均衡においては，他主体の最適な行動を前提にした，悲観的かつ安全側の行動を表現できる．

また，ライドシェア市場[6] や鉄道市場[7] における複数事業者の競合を分析する手法として，Stackelberg ゲームを適用する研究も見られる．このように，複数主体の相互動学的な意思決定を記述する手法として，交通計画分野で Stackelberg ゲームが用いられてきた．

しかし，Stackelberg ゲームは複数主体の再帰的な意思決定において，行動の読み合いが相互に起こり得ることを表現できていない．既往研究では，leader が follower の行動を予測できることを前提にした計画論を示しているが，follower もまた，leader の行動を予想して自らの行動を決定する場合がある．そのような場合，Stackelberg ゲームを多期間に拡張し，複数主体の意思決定の均衡を分析する必要がある．

(2) Security game と深層学習の適用

近年では，Stackelberg ゲームを，security game という警備計画の最適化に用いた研究が盛んである．Kiekintveld ら[3] は，悪意を持った攻撃者 (follower) が警備 (leader) の警備計画を観察できる前提で，leader の最適な警備計画を求めている．これは，攻撃者の最適行動を制約条件としてもつ警備側の最適化問題として記述することができ，ロサンゼルス国際空港の警備計画や国立公園内の密猟パトロールに実装されるなど注目を集めている．

Security game をネットワークに拡張した枠組みとして，Network Security Game (NSG) がある，NSG では，攻撃者 (follower) を捕まえるため警備員の位置を逐次的に操作する警備側 (leader) と，警備に捕まらないように目標ノードまで移動する攻撃者のゼロサムゲームを考える．これは，複数主体の先読み行動を考え，多期間にわたる選択の最適化を行っている点で，本研究と類似している．

しかし，大規模な NSG を解析的に解くのは一般に難しい．それは，leader と follower の行動選択肢が多く，ゲームが多期間に渡るため，先読みする状態数が膨大になるからである．これに対して，状態を入力とし最適な方策を出力する深層学習モデルを訓練することにより，近似的に最適戦略を導く研究がある．Xue ら[8] は，Neural Fictitious Self Play という手法を適用し，自己対戦 (self-play) を繰り返すことにより，leader と follower の最適反応の方策ネットワークを学習する枠組みを示した．また，Xue ら[9] は，囲碁 AI の一種である MuZero[10] の手法を用いて，モンテカルロシミュレーションにより教師データを生成し，大規模な NSG に対する defender の最適戦略を導いている．

(3) 本研究の位置付け

交通計画において，計画者と利用者の相互作用を記述するために Stackelberg ゲームが用いられてきた

が，利用者の先読み行動の記述と，多期間への拡張はなされていない．一方，Security game の分野では，近年，ゲームの状態空間の爆発に対処するため，深層学習の援用が試みられている．本研究では，計画者の社会基盤整備と住民の居住地選択の相互動学的な意思決定を記述するため，Stackelberg ゲームを多期間へと拡張した展開形ゲームの枠組みを提示する．また，ゲームの大規模化に対応するため，深層強化学習を適用し，均衡における計画者と住民の最適戦略を近似的に計算することを試みる．

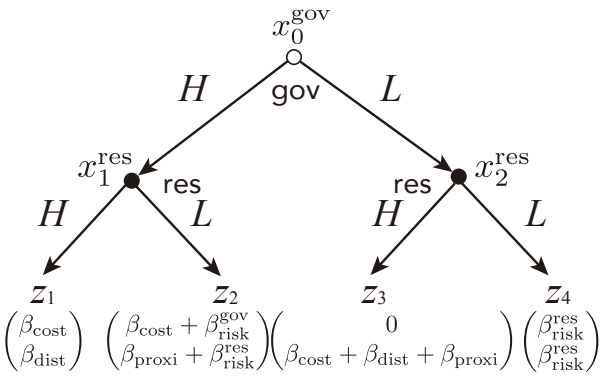


図-1　2 選択肢 1 期の展開形ゲーム (Stackelberg ゲーム)

3. 立地行動の展開形ゲームによる記述

本研究では，津波や水害等の災害が想定される災害リスクエリアにおける，行政の社会基盤投資と住民の居住地選択を扱う．災害リスクエリアにおいて，行政は，住民の被災リスクと土地の開発・維持管理コストを勘案しながら，既成市街地や未開発地における社会基盤投資計画を立てる．このとき，住民の被災リスクは住民の行動によって決まる．一方で，住民は，自らの被災リスクと居住地の利便性を考慮して，居住地を選択する．このとき，居住地の利便性は行政の計画に影響を受ける．このようにそれぞれの選択に互いの行動が影響を及ぼし合うため，先読みを伴う相互動学的な意思決定の表現に適した展開形ゲームを用いてモデル化する．

(1) 展開形ゲーム

本章では，計画者による社会基盤投資と住民による居住地選択の動学的な意思決定を展開形ゲームとして記述する．展開形ゲームの記述に用いる記号は以下の通りである．

- $I = \{\text{government, residents}\}$: ゲームのプレーヤー集合 I は，計画者 (government) と住民 (residents) から構成される．
- $x \in X$: 手番 (move)．ゲーム木上のノードで，後に起こり得るノードが存在するノード．X は手番の集合．
- x^{gov}: 計画者が行動する手番．
- x^{res}: 住民が行動する手番．
- s: 状態 (state)．各ゾーンについて計画者と住民の立地有無を 0/1 で表した大きさ 2 のベクトルを，ゾーンの数だけ並べたベクトル．
- $A(x^{\mathrm{gov,\,res}})$: 行動空間．それぞれ計画者と住民が各手番で取り得る行動の集合を表す．
- z: 頂点 (terminal nodes)．ゲーム木上のノードで，後に起こり得るノードが存在しないノード．プレーヤーが頂点 z に到達するとゲームは終了する．

各プレーヤーは，計画者から初めて順番に行動を繰り返す．計画者は社会基盤整備を行うゾーンを一つ選択し，住民は居住するゾーンを一つ選択する．本研究では，簡単のため計画者と住民の行動空間が同一であるとする．すなわち，l をゾーンを表す変数として，$A(x^{\mathrm{gov}}) = A(x^{\mathrm{res}}) = \{l \mid l = 1, 2, \ldots, L\}$ とする．計画者と住民がそれぞれ 1 度行動をとることを，1 期と呼ぶ．2 選択肢 1 期の場合の例を，**図-1** に示す．

頂点 z において，各プレーヤーは，自身及び他のプレーヤーの行動に基づく利得 (payoff) を得る．全プレーヤーの行動の結果は状態 s として記述されているため，利得は状態 s の関数である．計画者の利得 $u^{\mathrm{gov}}(s)$ は，土地の開発コストと住民の総被災リスクの線形和と仮定する．

$$u^{\mathrm{gov}}(s) = \beta_{\mathrm{cost}} \sum_l \overline{\delta_l^{\mathrm{new}}} \delta_l^{\mathrm{gov,1st}} + \beta_{\mathrm{risk}}^{\mathrm{gov}} \sum_l \overline{R_l} \delta_l^{\mathrm{res}} \quad (1)$$

β は各項の重みパラメータである．変数の上付き棒は，外生変数であることを表す．$\overline{\delta_l^{\mathrm{new}}}$ は，ゾーン l の新規開発にコストがかかる場合に 1，それ以外は 0 となる外生変数である．$\delta_l^{\mathrm{gov,1st}}$ は，計画者が住民の居住より先にゾーン l の社会基盤整備を行った場合に 1，それ以外は 0 となる二値変数である．つまり，式 (1) の第一項は，ゾーン l に新規開発コストが必要かつ計画者が住民より先にそのゾーンを開発した場合に，発生する土地の開発コストを表す．$\overline{R_l}$ は，ゾーン l の災害リスクを表す外生変数である．δ_l^{res} は，ゾーン l に住民が居住する場合に 1，それ以外は 0 となる二値変数である．よって，式 (1) の第二項は，各ゾーンのリスクの重み付き和，すなわち住民の総被災リスクを表す．

住民の利得 $u^{\mathrm{res}}(s)$ は，住民の総被災リスクと居住地の利便性の線形和と仮定し，居住地の利便性は，土地の開発コスト，既成市街地からの距離，インフラとの近接性，の線形和とする．

$$u^{\mathrm{res}}(s) = \beta^{\mathrm{res}}_{\mathrm{risk}} \sum_l \overline{R_l}\delta^{\mathrm{res}}_l + \beta_{\mathrm{cost}} \sum_l \overline{\delta^{\mathrm{new}}_l}\delta^{\mathrm{res,1st}}_l + \beta_{\mathrm{dist}} \sum_l \overline{D_l}\delta^{\mathrm{res}}_l + \beta_{\mathrm{proxi}} \sum_r \sum_g d_{r,g} \tag{2}$$

第一項，第二項は，式 (1) と同じく，それぞれ住民の総被災リスクと住民の土地開発コストを表す．第三項の $\overline{D_l}$ は，ゾーン l の中心市街地からの距離を表し，第三項全体では，住民が中心市街地から離れた場所に居住することのコストを表す．第四項の $d_{r,g}$ は，住民の選択した居住ゾーン r と計画者の選択した社会基盤整備ゾーン g の距離を表し，第四項全体では，計画者の選択と住民の選択の空間的乖離により住民が被るコストを表す．ゲームの均衡は，各プレーヤーの各局面における利得の大小に依存するため，5 つのパラメータ $\beta_{\mathrm{cost}}, \beta^{\mathrm{gov}}_{\mathrm{risk}}, \beta^{\mathrm{res}}_{\mathrm{risk}}, \beta_{\mathrm{dist}}, \beta_{\mathrm{proxi}}$ によって特徴づけられる．利得関数の全ての項は何らかのコストを表し，各項が増加すると利得は低下するため，5 つのパラメータは全て負である．

また，重要な仮定として，本研究では完備情報 (complete information) と完全情報 (perfect information) を仮定する．完備情報とは，全てのプレーヤーが，ゲームのプレーヤー集合や状態遷移, 各プレーヤーの行動空間・利得といったゲームのルールを保持していることを指す．完全情報とは，全てのプレーヤーが，自身が意思決定を行ういずれの手番においても，それ以前に行われた全てのプレーヤーによる意思決定の内容を観察できることを指す．実際には，計画者も住民も，不完備情報・不完全情報の下で意思決定を行っていると考えられるため，これらは強い仮定である．しかし，計画側から見ると，完全情報下で計画者の行動を先読みしながら合理的に意思決定を行う住民に対する，最適な行動を計算することは，悲観的で安全側に寄った計画を提示することを意味し，一定の意味を有すると考える．

(2) 部分ゲーム完全均衡

展開形ゲームの解は，部分ゲーム完全均衡という均衡概念を用いることにより導かれる．展開形ゲーム G の部分ゲームとは，根ノードが $x \in X$ であるゲーム木の一部で，その根ノードの子孫が元のゲーム木における根ノードの子孫と一致するようなものである．つまり，元のゲーム木の任意の手番とそれより下位のノードをそのまま切り出したゲーム木のことである．ゲーム G の部分ゲーム完全均衡とは，ゲーム G のいかなる部分ゲーム G_x においても，そこでのプレーヤーの戦略が x における最適反応となっている戦略の組 (strategy profile) のことである．つまり，部分ゲーム完全均衡は，元のゲームの全ての部分ゲームにおける純粋戦略ナッシュ均衡から構成されており，動学ゲームの均衡を表現するためにナッシュ均衡を精緻化した概念と言える．

部分ゲーム完全均衡を求めるための一般的な方法に，後向き帰納法がある．後向き帰納法では，まずゲームの最後の行動 (ゲーム木における頂点 z の直前のノード) から考える．最後に行動するプレーヤーが，あり得る全ての最終手番の中から，利得を最大化する行動を選択する．次に，最後に行動するプレーヤーの最適反応を前提にして，最後から二番目のプレーヤーが自らの利得を最大にする行動を決定する．これをゲームの最初の手番に至るまで繰り返すことにより，各手番におけるプレーヤーの最適反応を求めることができるため，部分ゲーム完全均衡が特定できる．プレーヤーは各手番において，その局面以降の自らと他のプレーヤーの合理的な行動を先読みして，再帰的に現在の行動を決定すると言える．ゲームの解は，最初の手番から各手番における最適反応を辿ることにより得られる．

4. 均衡解の性質

本章では，前章で述べた立地行動の動学ゲームの均衡解の性質を調べる．まず，2 選択肢 1 期の場合について，パラメータ条件と均衡解の関係を分析する．2 選択肢 1 期の場合は，2 プレーヤー (leader と follower) について，leader が follower の最適反応を見越して戦略を決定する Stackelberg ゲームである．次に，2 選択肢 3 期の場合について，パラメータ条件と均衡解の関係を分析する．多期間ゲームへと拡張することにより，Stackelberg ゲームと異なる均衡解が出現すると予想される．その理由は，leader (計画者) が follower (住民) の反応を予想するだけでなく，follower がさらに次期の leader (計画者) の行動を先読みして行動を決定し，leader がそれをもとに行動を決定する，という意思決定の相互性・動学性を表現できるためである．

(1) 2 選択肢 1 期の場合 (Stackelberg ゲーム)

災害リスクエリアにおいて，計画者 (leader) が，住民 (follower) の居住地選択を見越して，社会基盤整備場所を決定する Stackelberg ゲームを考える．計画者

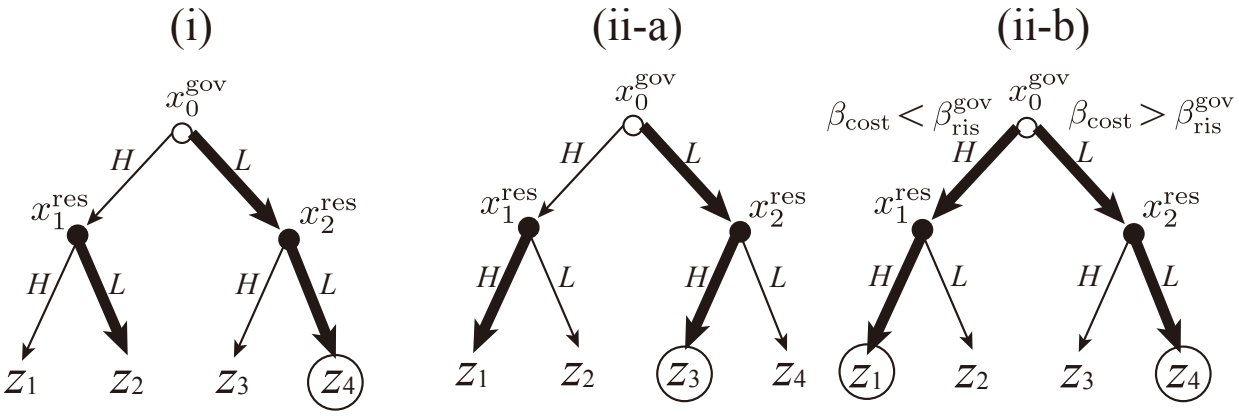


図-2　2 選択肢 1 期の展開形ゲームの部分ゲーム完全均衡

と住民の行動空間は共に高台 H か低地 L かの 2 選択肢とする．災害リスクエリアの一般的な状況として，低地 L に既成市街地が広がり，高台 H は未開発地である状況を想定する．高台 H のみ新規開発にコストが必要 ($\overline{\delta_H^{\mathrm{new}}} = 1, \overline{\delta_L^{\mathrm{new}}} = 0$) で，低地 L のみ災害リスクにさらされている ($\overline{R_H} = 0, \overline{R_L} = 1$)．既成市街地からの距離 $\overline{D_l}$ は，低地 L で 0，高台 H で 1 とする ($\overline{D_H} = 1, \overline{D_L} = 0, d_{H,H} = d_{L,L} = 0, d_{H,L} = d_{L,H} = 1$)．ゲーム木と，各頂点における計画者と住民の利得を**図-1** に示す．

ゲームの解をパラメータ条件ごとに整理する．ここで各パラメータ β は負であるため，パラメータが小さい方が絶対値が大きく，利得に与える影響度が大きいということに注意が必要である．パラメータ条件ごとの部分ゲーム完全均衡を**図-2** に示す．後向き帰納法を適用すると，まず x_1^{res} の手番において，**(i)(ii)** の 2 つの場合に分けられる．

(i) $\beta_{\mathrm{dist}} < \beta_{\mathrm{proxi}} + \beta_{\mathrm{risk}}^{\mathrm{res}}$ **のとき**

住民が，災害リスクやインフラとの近接性よりも，既成市街地への近さを重視する場合である．このとき，計画者が高台 H を選択した場合，x_1^{res} における住民の最適反応は低地 L を選択することである．その理由は，住民が高台 H を選択したときに発生する負の利得・コスト (β_{dist}) が，低地 L を選択したときに得るコスト ($\beta_{\mathrm{proxi}} + \beta_{\mathrm{risk}}^{\mathrm{res}}$) よりも，大きいからである．次に，計画者が低地 L を選択した場合，x_2^{res} における住民の最適反応は低地 L を選択することである．その理由は，条件から，$\beta_{\mathrm{risk}}^{\mathrm{res}} > \beta_{\mathrm{dist}} - \beta_{\mathrm{proxi}} > \beta_{\mathrm{dist}} + \beta_{\mathrm{proxi}} + \beta_{\mathrm{cost}}$ と導けるからである．

次に，住民の最適反応を前提にして，x_0^{gov} における計画者の行動を考える．計画者は，自らが高台 H を選択する場合，住民の最適反応が低地 L であると合理的に予測するため，計画者の利得は $\beta_{\mathrm{cost}} + \beta_{\mathrm{risk}}^{\mathrm{gov}}$ である．同様に，計画者が低地 L を選択する場合，計画者は住民の最適反応が低地 L であると合理的に予測するため，計画者の利得は $\beta_{\mathrm{risk}}^{\mathrm{gov}}$ である．$\beta_{\mathrm{cost}} + \beta_{\mathrm{risk}}^{\mathrm{gov}} < \beta_{\mathrm{risk}}^{\mathrm{gov}}$ であるため，x_0^{gov} での計画者の最適反応は，低地 L を選択することである．

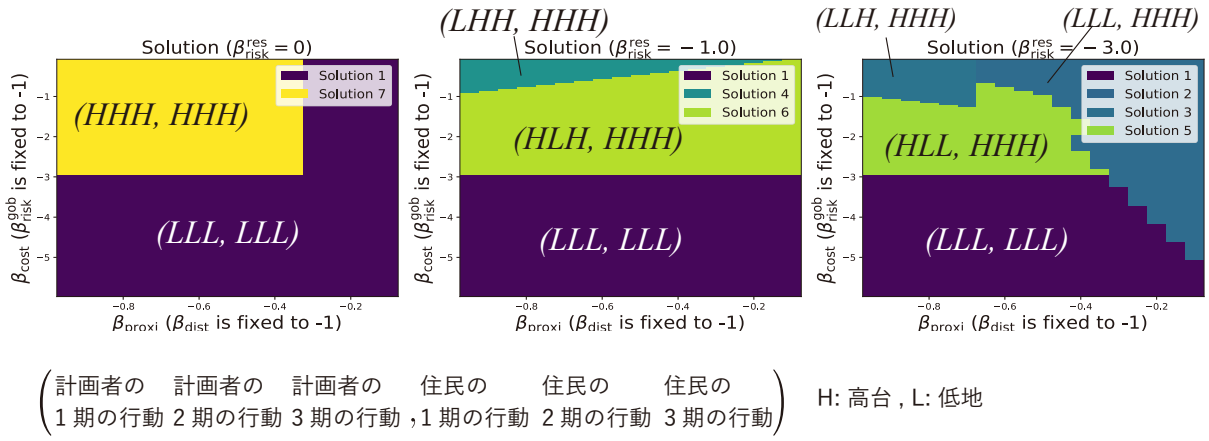
以上のことから，住民が災害リスクやインフラとの近接性よりも，既成市街地への近さを重視する場合，計画者のリスク認知や高台の開発コストにかかわらず，計画者も住民も低地 L を選択することが均衡解となる．計画者は，計画者の選択に関係なく住民が低地 L を選択することを合理的に予測するため，高台 H の開発コストを無駄に払わないように，低地 L を選択するためである．

(ii) $\beta_{\mathrm{dist}} > \beta_{\mathrm{proxi}} + \beta_{\mathrm{risk}}^{\mathrm{res}}$ **のとき**

住民が，既成市街地への近さよりも，災害リスクやインフラとの近接性を重視する場合である．このとき，計画者が高台 H を選択した場合，x_1^{res} における住民の最適反応は高台 H を選択することである．次に，計画者が低地 L を選択した場合，x_2^{res} における住民の最適反応はさらに 2 つの場合に分けられる．

(ii-a) $\beta_{\mathrm{cost}} + \beta_{\mathrm{dist}} + \beta_{\mathrm{proxi}} > \beta_{\mathrm{risk}}^{\mathrm{res}}$ **のとき**

住民のリスク認知が，高台の開発と既成市街地からの距離とインフラとの近接性によるコストよりも，さらに大きい場合である．このとき，計画者が低地 L を選択した場合，x_2^{res} における住民の最適反応は高台 H を選択することである．x_0^{gov} において，計画者が高台 H を選択した場合，住民は高台 H を選択するため，計画者の利得は β_{cost} である．計画者が低地 L を選択した場合，住民は高台 H を選択するため，計画者の利得は 0 である．計画者は住民の最適反応を先読みするため，低地 L を選択し，住民が高台 H を選択することが均衡解となる．つまり，住民のリスク認知が十分に大きい場合，計画者は自らの選択に関係なく住民が高台 H を選択することを合理的に予測するため，高台 H の開発コストを無駄に払わない



$\left(\begin{matrix}\text{計画者の}\\ \text{1 期の行動}\end{matrix}\ \begin{matrix}\text{計画者の}\\ \text{2 期の行動}\end{matrix}\ \begin{matrix}\text{計画者の}\\ \text{3 期の行動}\end{matrix}\ ,\begin{matrix}\text{住民の}\\ \text{1 期の行動}\end{matrix}\ \begin{matrix}\text{住民の}\\ \text{2 期の行動}\end{matrix}\ \begin{matrix}\text{住民の}\\ \text{3 期の行動}\end{matrix}\right)$ H: 高台 , L: 低地

図-3　2 選択肢 3 期の展開形ゲームの部分ゲーム完全均衡

ように，低地 L を選択する．

(ii-b) $\beta_{cost}+\beta_{dist}+\beta_{proxi}<\beta_{risk}^{res}$ **のとき**

住民のリスク認知が，高台の開発，既成市街地からの距離，インフラとの近接性によるコストよりは，小さい場合である．このとき，計画者が低地 L を選択した場合，x_2^{res} における住民の最適反応は低地 L を選択することである．x_0^{gov} において，計画者が高台 H を選択した場合，住民は高台 H を選択するため，計画者の利得は β_{cost} である．計画者が低地 L を選択した場合，住民は低地 L を選択するため，計画者の利得は β_{risk}^{res} である．よって，β_{cost} と β_{risk}^{res} の大小により，さらに 2 つの場合に分けられる．

$\beta_{cost}<\beta_{risk}^{res}$，つまり高台の開発コストが計画者のリスク認知より影響が大きい場合，計画者は x_0^{gov} において低地 L を選択するため，計画者と住民の双方が低地 L を選択することが均衡解となる．$\beta_{cost}>\beta_{risk}^{res}$，つまり計画者のリスク認知が高台の開発コストより大きい場合，計画者は x_0^{gov} において高台 H を選択するため，計画者と住民の双方が高台 H を選択することが均衡解となる．

つまり，住民のリスク認知がある程度大きいが，高台の開発，既成市街地からの距離，インフラとの近接性によるコストよりは小さい場合，計画者・住民とも高台を選択するか，反対に計画者・住民とも低地を選択する．これらの解の違いは，計画者側の決定に依存する．

(2)　2 選択肢 3 期の場合

前節で述べた 2 選択肢 1 期の Stackelberg ゲームを，2 選択肢 3 期の多期間の動学ゲームへと拡張する．2 選択肢 3 期の場合，ゲーム木の頂点 z の数が，$2^{2\times3}=256$ 通り存在し，前節のようにパラメータの場合分けで均衡解を分析することは困難である．そこで，さまざまなパラメータの組み合わせのもと，後向き帰納法により部分ゲーム完全均衡における解を求め，解空間を分析する．外生変数 $\overline{\delta_H^{new}}=1,\overline{\delta_L^{new}}=0,\overline{R_H}=0,\overline{R_L}=1,\overline{D_H}=1,\overline{D_L}=0,d_{H,H}=d_{L,L}=0,d_{H,L}=d_{L,H}=1$ の設定は前節と同様である．

まず，均衡解のバラエティについて調べるため，5 つのパラメータをそれぞれ 5 段階で変化させ，$5^5=3125$ 通りのパラメータの組について，ゲームの解を求めた．各パラメータのとる数値は以下の通りである．

- β_{cost}：$-0.3,-1.5,-3.0,-4.5,-9.0$
- $\beta_{risk}^{gov,res},\beta_{dist}$：$-0.1,-0.5,-1.0,-1.5,-3.0$
- β_{proxi}：$-0.03,-0.15,-0.3,-0.5,-1.0$

式 (1)，式 (2) において，パラメータがかかっている各数値の大きさに違いがあるため，利得に与える影響を標準化する必要がある．そのため，パラメータの絶対的な大きさが異なっている．

計算の結果，均衡解として 7 種類の解が現れることがわかった．計画者の各期の行動を $a_{1,2,3}^{gov}$，住民の各期の行動を $a_{1,2,3}^{res}$ と表記し，ゲームの解を $(a_1^{gov}a_2^{gov}a_3^{gov},a_1^{res}a_2^{res}a_3^{res})$ と表記する．このとき，7 種類の解は以下の通りである．

1. (LLL,LLL): 計画者，住民ともに低地を選択．
2. (LLL,HHH): 計画者のみ低地を選択し，住民は高台を選択．
3. (LLH,HHH): 計画者が途中で高台に選択を変更．住民は高台を選択．
4. (LHH,HHH):計画者が途中で高台に選択を変更．住民は高台を選択．
5. (HLL,HHH): 計画者が途中で低地に選択を変

更．住民は高台を選択．

6. (HLH, HHH): 計画者が低地に選択を変更し，その後再び高台を選択．住民は高台を選択．
7. (HHH, HHH): 計画者，住民ともに高台を選択．

図-3 に，パラメータの値とそれぞれの解の存在範囲を図示する．なお，利得の大きさの絶対的な水準を固定するため，$\beta_{\text{risk}}^{\text{gov}} = -1, \beta_{\text{dist}} = -1$ と固定し，$\beta_{\text{cost}}, \beta_{\text{risk}}^{\text{res}}, \beta_{\text{proxi}}$ について感度分析を行った．**図-3** の 3 枚の図は，それぞれ $\beta_{\text{risk}}^{\text{res}} = 0, -1.0, -3.0$ の場合に，β_{cost} と β_{proxi} を動かしたときの，均衡解の変化を示した図である．

まず，$\beta_{\text{risk}}^{\text{res}} = 0$，つまり住民が災害リスクを無視する場合を考える．このとき，ほとんど全ての領域で解 $1(LLL, LLL)$ となる．一方で，β_{cost} の絶対値が小さく，β_{proxi} の絶対値が大きい領域では，解 $7(HHH, HHH)$ が現れる．つまり，住民が居住地選択において災害リスクを考慮しないとしても，高台の新規開発コストが小さく，住民がインフラとの近接性を重視する場合には，災害リスクの高い低地部から高台への移転が起こり得る．

次に，$\beta_{\text{risk}}^{\text{res}} = -1$，つまり住民の災害リスクが中程度である場合を考える．β_{cost} の絶対値が大きい，つまり高台の新規開発コストが大きい場合は，住民のインフラとの近接性への選好に関わらず，解 $1(LLL, LLL)$ となる．β_{cost} の絶対値が小さくなると，解 $6(HLH, HHH)$ が現れる．このとき，計画者が先行して高台 H に投資することにより，住民が高台を選択することを先読みするため，新規開発コスト β_{cost} を負担する．さらに高台開発コストが低下すると，解 $4(LHH, HHH)$ が現れる．このとき，新規高台開発コスト β_{cost} が非常に小さいため，計画者は，住民が被災リスクを低下させるために自ら高台を選択することを先読みする．そのため，1 期目の選択では低地 L を選択し，新規高台開発コストを住民に負担させる．

最後に，$\beta_{\text{risk}}^{\text{res}} = -3.0$，つまり住民の災害リスク認知が非常に高い場合である．新規高台開発コスト β_{cost} の絶対値が大きいときは，これまでと同様に，解 $1(LLL, LLL)$ となり，低地に居住と開発が集中する．一方，新規高台開発コスト β_{cost} の絶対値が小さいと，解 $2(LLL, HHH)$ が現れる．計画者は低地を選び続け，住民は高台を選び続ける解である．住民が居住地選択において災害リスクを重視し，自ら新規高台開発コストを負担することを，計画者は合理的に予測するため，高台の社会基盤整備を行わない．また，β_{proxi} の絶対値が大きいとき，解 $5(HLL, HHH)$ が現れる．これは，住民が計画者の整備したインフラに近い場所を選択することを重視することを，計画者が先読みし，住民に先行して高台 H の社会基盤整備を行うことを意味している．

以上のように，計画者と住民の立地行動を，多期間の動学ゲームへと拡張した場合，途中での行動変更がゲームの解になり得る．これは，新規高台開発コストが，高台を先に開発したプレーヤーにしか発生せず，当該行動をとるタイミングが重要であるためと考えられる．多期間の展開形ゲームは，本章の新規高台開発コストのように，複数主体が特定の行動をとる順序や経過時間が意思決定に影響する場合の記述に優れているといえる．

5. マルチエージェント強化学習による大規模計算

(1) AlphaZero

展開形ゲームの解を求める際，後向き帰納法による全探索の計算コストは，プレーヤーの行動空間の大きさと，期間の数に応じて指数的に増大する．4 章の例で，計画者と居住者の行動選択肢の数をどちらも L とし，T 期の意思決定を考える場合，後向き帰納法では L^{2T} の頂点を評価する必要がある．本章で扱う 6 選択肢 10 期の展開形ゲームでは，頂点数は $6^{10} \approx 10^7$ であり，ナイーブな後向き帰納法により解を求めることは困難である．そこで，本章では，マルチエージェント深層強化学習の AlphaZero[11] を用いて，ゲームの解を近似的に求める方法を提示する．

AlphaZero は，囲碁や将棋などのボードゲーム AI の一種で，人間のプロ棋士に勝利した AlphaGo の後継モデルである．人間の知識や過去の対戦データを必要とせずに，複数エージェント同士の自己対戦 (self-play) のみから最適戦略を計算することに特徴がある．**図-4** に，AlphaZero による最適戦略の探索のフレームワークを示す．立地行動の展開形ゲームに適用するため，ボードゲームの盤面をゾーンに区切られた都市空間と捉える．計画者と住民は，各局面において，その後の相手の戦略を先読みしながら，インフラと居住地の最適な配置ゾーンを決定する．AlphaZero は，盤面 (都市空間) の状態を入力とし，方策と報酬を出力する deep neural network により，筋の悪い行動を足切りし，効率的に最適な戦略を探索する．Deep neural network は，自己対戦の繰り返しによって得た，盤面・その時の方策・最終的な報酬によって訓練される．また，自己対戦では，モンテカルロ木探索により，各局面から最終的な報酬を最大化する行動を探索する．

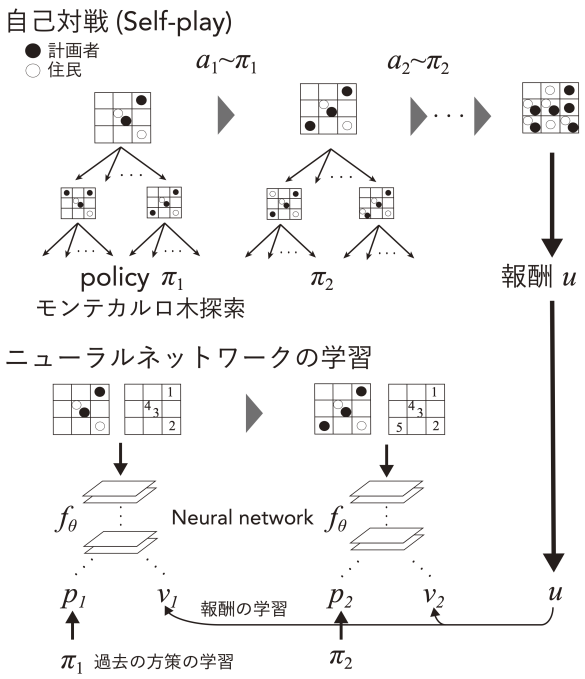


図-4　AlphaZero のフレームワーク

a)　モンテカルロ木探索

AlphaZero は，モンテカルロ木探索に，UCT (Upper Confidence bound applied to Trees) アルゴリズムを適用している．UCT アルゴリズムは，多腕バンディット問題のアルゴリズムで，限られた回数のプレイアウトから，現在の盤面に対する最適な行動を探索するために使用する．ここで，プレイアウト (play-out) とは，現在の盤面からゲームの終了までランダムな対戦を行うことを言う．本研究では，自己対戦中に出現する全ての盤面に対して，ゲームの終端まで 50 回のプレイアウトを行った．プレイアウトにおいて，AlphaZero は，式 (4), (5) で計算される UCT スコアが最大の行動を選択する (式 (3))．これにより，過去の経験の活用と，未知の行動の探索のバランスをとりながら，最適戦略を探索することが可能となる．

$$a_t = \underset{a}{\mathrm{argmax}}(Q(s_t,a) + U(s_t,a)) \tag{3}$$

$$Q(s_t,a_t) = \frac{W(s_t,a_t)}{N(s_t,a_t)} \tag{4}$$

$$U(s_t,a_t) = \eta P(s,a)\frac{\sqrt{N(s_t)}}{1+N(s_t,a_t)} \tag{5}$$

$Q(s_t,a)$ は，盤面の状態 s_t で行動 a が取られた場合の勝率であり，過去のプレイアウトの経験を表す．添字の t は，現時点までのプレイアウトの繰り返し回数である．式 (4) の $W(s_t,a_t)$ は，状態 s_t で行動 a_t をとる場合の累積報酬である．$N(s_t,a_t)$ は，状態 s_t で行動 a_t をとったプレイアウトの回数である．盤面を入力とし，その盤面が最終的に勝つか負けるかを予測するニューラルネットワークの出力を v とする．AlphaZero では，累積報酬 $W(s_t,a_t)$ を更新するため，ニューラルネットワークによる盤面評価値 v を用いる．

$U(s,a)$ は探索を表す項で，探索回数 $N(s_t,a_t)$ が大きいほど小さい値をとる．パラメータ η は探索と活用のバランスを表す．本研究では，$\eta = 1$ とした．$P(s,a)$ は，盤面状態 s で行動 a をとることの事前確率の役割を果たしている．AlphaZero では，$P(s,a)$ は，盤面を入力としこれまでの行動方策を出力するニューラルネットワークから得られる．過去の探索データに基づいて，勝利する確率の低い行動の UCT スコアを減じることにより，探索の効率性を向上させている．

b)　Deep neural network

前項で述べたように，AlphaZero ではモンテカルロ木探索において，$(\boldsymbol{p}, v) = f_\theta(s)$ で表現されるニューラルネットワークを用いる．θ はニューラルネットワークのパラメータである．ニューラルネットワークは，盤面の状態 s を入力とし，計画者と住民の立地確率ベクトル $\boldsymbol{p}$ と期待報酬 v を出力する．本研究では，畳み込みニューラルネットワーク (CNN) を用いた．ネットワークのアーキテクチャを**表-1** に示す．CNN の入力は，都市空間を表す 6×6 のグリッドで，4 つのチャネルをもつ．一つ目は計画者の社会基盤整備場所，二つ目は住民の立地場所，三つ目は社会基盤整備からの経過時間，四つ目は住民の立地からの経過時間である．経過時間のチャネルを使用した理由は，計画者と住民のどちらが先に投資するかにより，式 (1)(2) 内の土地の新規開発コストが異なるためである．学習の安定化のため，畳み込み層と全結合層の後にバッチ正規化層を挿入する．立地確率ベクトル $\boldsymbol{p}$ と期待報酬 v を出力するネットワークは全結合層以外は同じパラメータを共有する．損失関数は，予測報酬の二乗誤差と予測方策のクロスエントロピーの二指標の平均である．最適化は Adam で行い，学習率は 0.0005 とした．

(2)　結果と考察

都市空間を模した 6×6 のグリッド上で AlphaZero の学習とシミュレーションを行った．まずシミュレーションの設定を示す．式 (2) の $\overline{D_l}$ の設定として，最も左上のゾーンを既成市街地とした．リスクは上のグリッドほど高く，下のグリッドほど低いと設定した．ゲームは 10 期まで行う．利得の各パラメータは次のように設定した: $\beta^{\mathrm{gov}}_{\mathrm{cost}} = -9.0, \beta^{\mathrm{gov}}_{\mathrm{risk}} = -1.0, \beta^{\mathrm{res}}_{\mathrm{cost}} = -3.0, \beta^{\mathrm{res}}_{\mathrm{risk}} = -0.3, \beta_{\mathrm{dist}} = -1.0, \beta_{\mathrm{proxi}} = -0.03$.

表-1　畳み込みニューラルネットワークのアーキテクチャ

層	フィルター数	カーネルサイズ	Padding	活性化関数	出力
Input	3				(6,6,3)
Convolution	256	3 × 3	True	ReLU	(6,6,256)
Convolution	256	3 × 3	True	ReLU	(6,6,256)
Convolution	256	3 × 3	True	ReLU	(6,6,256)
Convolution	256	3 × 3	True	ReLU	(6,6,256)
Flatten					9216
Fully Connected	1024			ReLU	1024
Dropout	0.25				
Fully Connected	256			ReLU	256
Dropout	0.25				
Fully Connected (policy)	37			Softmax	37
Fully Connected (value)	1			Tanh	1

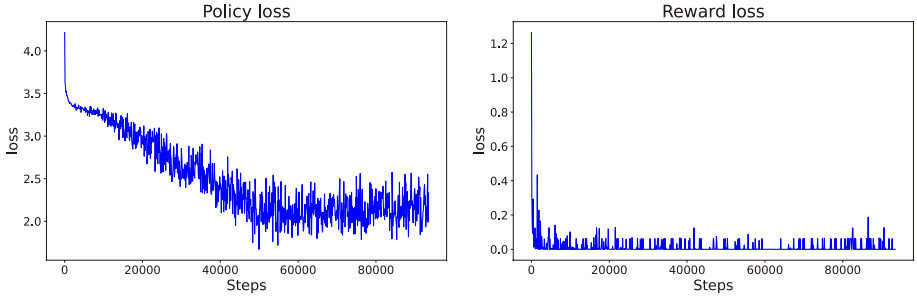


図-5　方策と報酬の損失関数の推移

自己対戦を 4000 回繰り返し，100 回行われるごとに，自己対戦における各盤面・報酬・方策を教師データとして畳み込みニューラルネットワークを学習する．また，自己対戦が 100 回行われるごとに，50 回のテストプレイを行い，AlphaZero の戦略と獲得利得を確認する．テストプレイでは，AlphaZero が計画者か住民かをランダムに選択し，モンテカルロ木探索に基づいて戦略を決定する．一方，AlphaZero でない方のプレーヤーは，貪欲・近視眼的な戦略を採用する．つまり，ε の確率で立地場所をランダムに選択し，$1-\varepsilon$ の確率で次期の報酬を最大にする立地場所を選択する．パラメータ $\varepsilon = 0.3$ とする．ゲームの報酬は，計画者と住民それぞれについて平均的な利得の値に対する獲得利得の増加割合を計算し，増加割合の大きい方が最終的な報酬 1 を獲得し，小さい方は報酬-1 とした．パラメータ値や変数によって計画者と住民で利得のスケールが異なるため，そのまま大小比較を行えないからである．自己対戦を始める前に，貪欲な戦略をお互いにとる場合の計画者と住民の平均利得を式 (1), (2) を使って計算しておく．

図-5 左は，AlphaZero の畳み込みニューラルネットワークによる予測方策と真の方策のクロスエントロピー誤差を表し，図-5 右は予測報酬と真の報酬の事情誤差を表す．どちらも自己対戦を重ねるにつれて，学習が進み誤差が低減している．報酬の誤差は，報酬の誤差に比べて非常に早く誤差が低減しており，学習速度が速い．

次に，図-6 は，学習前後の AlphaZero の戦略の変化を示した図である．左列は自己対戦が 100 回目，つまり AlphaZero の学習前のテストプレイで，右列は自己対戦が 4000 回目，つまり学習後のテストプレイである．上段は，計画者が AlphaZero の場合，下段は，住民が AlphaZero の場合である．グリッド中の青四角は計画者がそのゾーンに社会基盤整備を行ったことを表し，黄四角は住民が立地したことを表す．数字は，立地したタイミングである．

まず，上段に着目すると，AlphaZero によって最適反応を学習した計画者は，リスクの低い土地に社会基盤整備を行う．また，計画者がリスクの低い土地を開発すると，住民は β_{proxi} の効果によりその周辺に立地することの利得が高まる．そのため，自己対戦が 4000 回目のとき，社会基盤の整備されたリスク

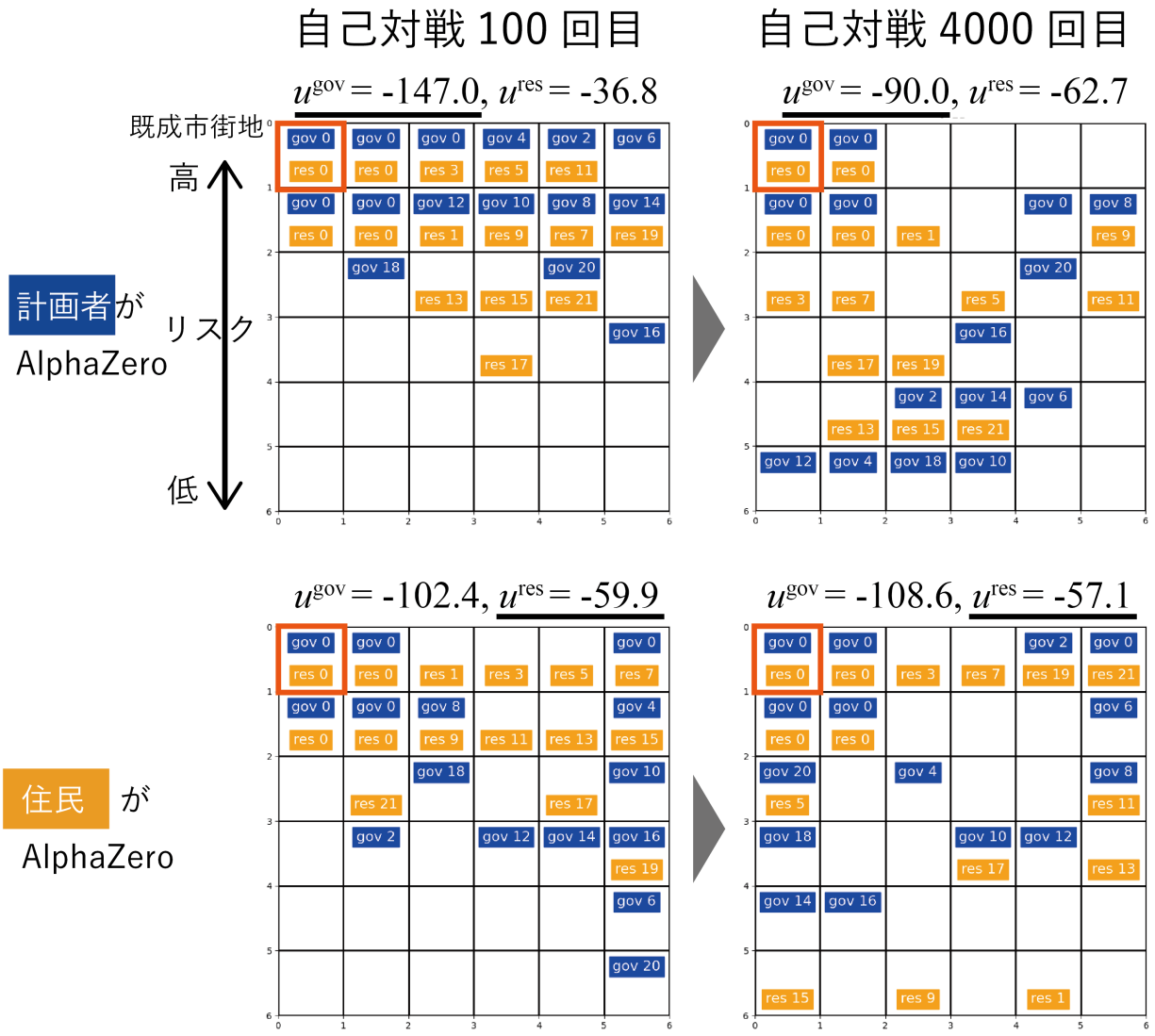


図-6　学習前と学習後の AlphaZero の戦略の変化

の低い土地に住民の立地が引き出されている．2 つのテストプレイにおいて，計画者の利得を比較すると，-147.0 から-90.0 と利得が大きく向上した．つまり，AlphaZero によって計画者は自らの利得を増加させる戦略を学習した．

次に，下段に着目すると，AlphaZero によって最適反応を学習した住民は，リスクの高い土地に立地を続けながらも，その一部はリスクの低い土地へ移転している．リスクの低い土地に立地したタイミングは，多くの場合，周辺の社会基盤整備に先行している．よって，住民は，計画に追従するだけではなく，自らの被災リスクを低下させるため一部がリスクの低い土地へ移転する．2 つのテストプレイにおいて，住民の利得を比較すると，-59.9 から-57.1 と利得が向上した．学習前後で，全テストプレイの平均利得を比較すると，計画者が AlphaZero の場合の計画者の平均利得は-151.6 から-125.6 へと大きく向上した．一方で，住民が AlphaZero の場合の住民の平均利得は-64.3 から-64.0 へと微増に止まった．住民は既成市街地の利便性とリスクの高さのトレードオフに直面しており，利得を向上させ辛いゲーム構造である可能性が考えられる．

以上より，AlphaZero を立地行動の動学ゲームに適用することにより，プレーヤーは自らの利得を増やす行動戦略の学習に成功した．学習が完全に行われた場合，自己対戦において，あるプレーヤーは他方のプレーヤーの最適反応をモンテカルロ木探索により先読みしながら，現在の最適な行動を選択するため，部分ゲーム完全均衡を求めていることになる．計画者の側から見ると，AlphaZero による社会基盤整備計画は，住民の合理性と先読みを前提とした上での最適な投資戦略を示している．意思決定の相互性

による，住民と計画者の選択の多様な展開を予測した上で，最適な政策を提示するため，将来を見据えた効果的な政策を打つことができると期待される．

6. 結論

本研究では，社会基盤・交通分野における多主体の相互動学的な意思決定をモデル化し，その均衡を分析するフレームワークを提示した．災害リスクエリアにおける，計画者の社会基盤整備と住民の居住地選択の再帰的な意思決定を対象に，まず，多主体の読み合いを展開形ゲームで記述し，部分ゲーム完全均衡を求めることにより，ゲームの解を特定した．計画者の投資と住民の立地が低地に集中する場合，高台に集中する場合，その中間等の解のバリエーションが，利得関数のパラメータの変化によって出現することを明らかにした．また，大規模なゲームでは状態空間の爆発により，各プレーヤーの各状態での最適反応をナイーブに求めることが困難になる．これに対し，マルチエージェント深層強化学習の一種である AlphaZero を用いて，各状態の最適反応を近似的に求めた．AlphaZero による最適な行動戦略の下で，計画者の利得は大きく向上し，多主体の相互作用と将来の展開を先読みした行動戦略が社会基盤整備計画に有効である可能性が示唆された．

最後に今後の研究の方向性を展望する．まず，動学的立地ゲームの不完全情報ゲームへの拡張が考えられる．本研究では，計画者と住民が完備情報・完全情報の下で合理的に意思決定を行うと仮定してモデリングを行った．しかし，実際には両者とも相手の行動を全て観察できるわけではないため，不完全情報の仮定が自然である．これにより，他者の行動の不確実性を記述し，不確実性下での最適戦略と完全情報下での最適戦略の差を計算できる．近年研究が進められている，機械学習や大規模言語モデルの手法を用いて，不完全情報ゲームにおけるナッシュ均衡を求めることが一つの方向性として考えられる．また，住民の合理性の仮定を緩和し，限定合理性や先読みの期数の制限を導入することにより，均衡解と最適な戦略のロバスト性を検討することも今後の課題である．最後に，複数主体の協力ゲームへの拡張が考えられる．本研究は立地行動の動学ゲームをゼロサムゲームとして記述したが，計画者と住民の協力行動下での均衡を検討することも重要である．計画者がコミットメント等の情報戦略をとる場合や協力ゲームの均衡を考えることにより，望ましい均衡への誘導を検討することは今後の課題である．

謝辞：本研究は JSPS 科研費 JP23KJ0771 の助成を受けたものです．ここに感謝の意を表します．

REFERENCES

1) 朝倉康夫. 利用者均衡を制約とする交通ネットワークの最適計画モデル. **土木計画学研究・論文集**, 6:1–19, 1988.
2) Miguel F Anjos, Bernard Gendron, and Martim Joyce-Moniz. Increasing electric vehicle adoption through the optimal deployment of fast-charging stations for local and long-distance travel. *European journal of operational research*, 285(1):263–278, 2020.
3) Christopher Kiekintveld, Manish Jain, Jason Tsai, James Pita, Fernando Ordónez, and Milind Tambe. Computing optimal randomized resource allocations for massive security games. 2009.
4) James Pita, Manish Jain, Milind Tambe, Fernando Ordónez, and Sarit Kraus. Robust solutions to stackelberg games: Addressing bounded rationality and limited observations in human cognition. *Artificial Intelligence*, 174(15):1142–1171, 2010.
5) Michael GH Bell and Chris Cassir. Risk-averse user equilibrium traffic assignment: an application of game theory. *Transportation Research Part B: Methodological*, 36(8):671–681, 2002.
6) Zeen Cai, Dong Mo, Wei Tang, Yong Chen, and Xiqun Michael Chen. A two-period game-theoretical model for heterogeneous ride-sourcing platforms with asymmetric competition and mixed fleets. *Transportation Research Part E: Logistics and Transportation Review*, 178:103279, 2023.
7) Dongjun Li, Dewan Md Zahurul Islam, Mark Robinson, Dong-Ping Song, Jing-Xin Dong, and Marc Reimann. Network revenue management game in the railway industry: Stackelberg equilibrium, global optimality, and mechanism design. *European Journal of Operational Research*, 312(1):240–254, 2024.
8) Wanqi Xue, Youzhi Zhang, Shuxin Li, Xinrun Wang, Bo An, and Chai Kiat Yeo. Solving large-scale extensive-form network security games via neural fictitious self-play. *arXiv preprint arXiv:2106.00897*, 2021.
9) Wanqi Xue, Bo An, and Chai Kiat Yeo. Nsgzero: Efficiently learning non-exploitable policy in large-scale network security games with neural monte carlo tree search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4646–4653, 2022.
10) Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.
11) David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

(Received ??? ?? 202?)
(Accepted ??? ?? 202?)

MUTUAL DYNAMIC DECISION-MAKING OF LOCATION CHOICE AND PROPERTY OF THE SOLUTION

Satoki MASUDA and Eiji HATO